Novel Applications of Machine Learning and Modelling in CMP Process Control To Minimize CMP Process Marginalities

John Matovu, *Principal Engineer*, OCT ADT, Cynthia Atanga, *Data Scientist*, Giovanni Mazzone, *Principal Integration Engineer*, Kevin Gast, *Sr CMP Engineer.*

Abstract: Chemical Mechanical Planarization (CMP) has challenges in process control resulting in uniformity and topography problems that culminate into downstream process marginalities and yield fails. The current control schemes normally use pre thickness and tool polish parameters such as polish time to control the process. In this work, we propose using several incoming step parameters together with inline CMP polish parameters to develop models to control the process. As an example, a poly CMP stop on oxide has been used to develop models using DELTA OXIDE THK as the control metric. During model development, the analysis shows that upstream stress factors such as Bow and Warp in the Pre Photo Shape contribute to uniformity variation. Therefore, building a controller that uses a combination of both upstream steps and the inline CMP polish process parameters could possibly enhance the ability to control the DELTA OXIDE variation better. Exploratory and predictive analysis on the upstream step data is done using some statistical analysis and machine learning algorithms to identify a possible good controller from the upstream step parameters. Promising results are shown from model building with these data. The next steps will require combining this model with a process step controller to amplify predictive power of DELTA OXIDE THK.

Index Terms: **CMP, DELTA, RPA.**

I. INTRODUCTION

Today several CMP processes have challenges related to process control leading to uniformity and topography issues that eventually lead to probe fails. For example, for CMP processes with very thick oxide stacks of several microns, incoming stress and other factors have been shown to impact outgoing oxide uniformity thickness. These processes normally have Integrated Automated Process Control (iAPC) which uses only incoming pre CMP thickness but does not include incoming topography and stress. Figures 1 & 2 illustrate these issues.

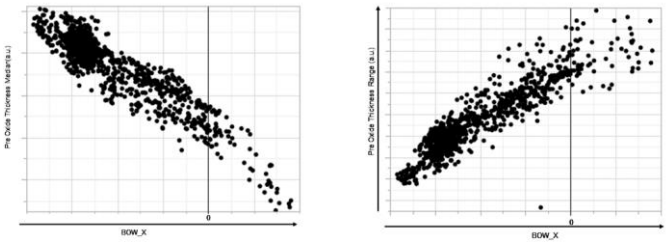

Figure 1: Graph showing Median and Range pre CMP thickness versus x-Bow

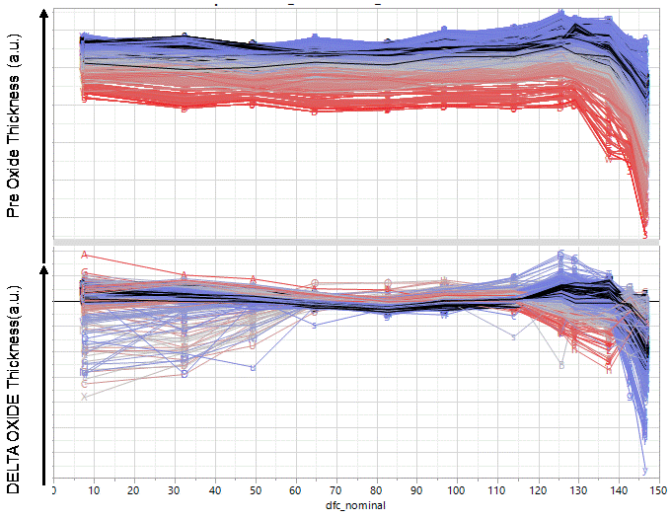

Figure 2: Graph showing PRE and DELTA CMP thickness versus DFC (Distance From Middle)) for a thick Oxide CMP Process

These similar problems shown in Figure 3 are being experienced for other process steps.

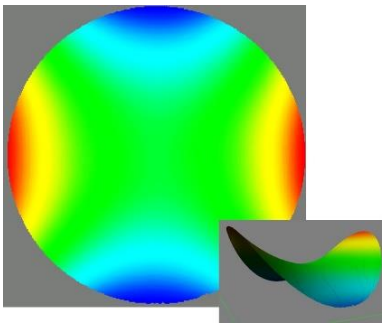

Figure 3: Pre CMP PWG Image showing Stress on the Wafer for a thick Oxide CMP Process.

At Poly CMP, for example, incoming factors such as incoming topography, and stress seem to impact the overall DELTA OXIDE removal thickness uniformity. Figure 4 shows a graphical correlation of DELTA OXIDE to incoming topography.

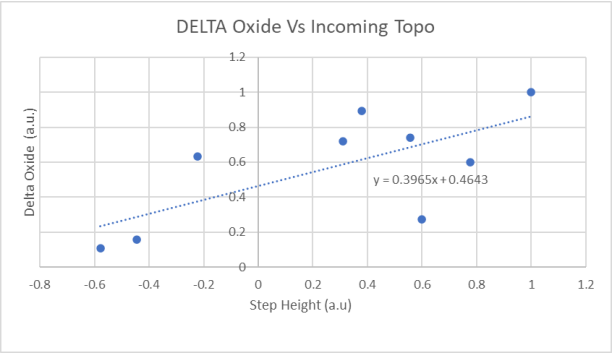

Figure 4: Graph shows Impact of Incoming Step topography to Oxide Delta for Poly CMP

The current R2R control schemes use only 1 single pre CMP (Figure 5) thickness and do not incorporate incoming topography and stress.

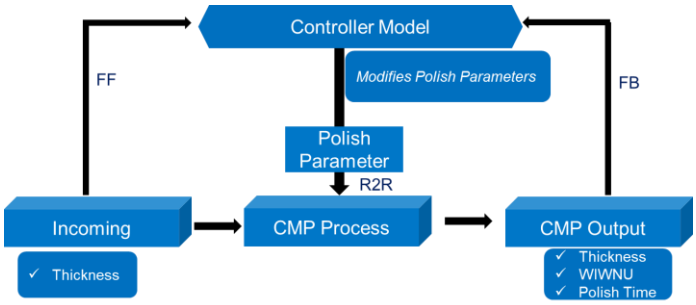

Figure 5: Conventional Area Process Control Scheme

Machine Learning in process control has been mostly focused on Area Process Control (APC) to enable better process control (1, 2). A few studies have shown modelling and machine learning considering a few steps in a module. For example, R. Ghulghazaryan and J. Wilson modelled various oxide deposition profiles on the impact of material removal rates during CMP (3). These still fall short of modelling and showing the impact of upstream steps on variations that are observed during the CMP process.
In this work, we focus on Module Process Control (MDC) for the machine learning Model schemes (Figure 6) that employ module-based data to control CMP processes. In this work, modelling for this poly module has been used to develop models for better process control.

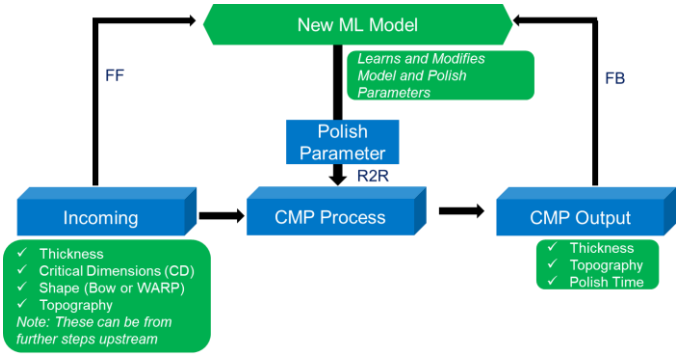

Figure 6: New Module Process Control Scheme

The primary goal is to assess the factors upstream and within the process that influence the DELTA OXIDE THK of wafers in the POLY CMP step, and then use them to predict the DELTA OXIDE THK. This process is broken down into two steps: upstream step modeling and process step (Recipe Parameter Adjustment (RPA)) modeling. This paper will focus on the former.
To our knowledge, no module machine learning and modeling work of CMP processes has been recently published. However, recently, M. Meiners and J. Franke used Machine Learning to apply Cross-Machine Control Loop to a downstream process from upstream processes in large series manufacturing (4). Although, their work is not in CMP, it provides useful insights in modeling that we use in this work. Their approach involved identifying relevant parameters upstream that showed high linear correlation to the downstream parameters, as well as multivariate modeling where linear relationships could not be easily identified. A ML regression algorithm was used to map the inter-relationships between these parameters and the downstream response as the complexities between various processes couldn't be identified by linear correlations.

Modeling with the upstream steps will allow prediction of the response (Poly CMP DELTA_OXIDE_THK), providing feedback to the process, while an informed decision based on predictions from process parameters can help guide on which of the process parameters to be adjusted to produce optimal output. Of course, not all process parameters can be adjusted, hence subject matter expert (SME) knowledge on the process parameters to be adjusted will be needed in deciding the parameters to be used in the model.

In modeling for the upstream parameters, current analysis involves multivariate modeling, where overlapping steps are combined and used to build the model. The steps were picked after evaluation of all steps in the module, based on sample size and correlation to response. In the analysis, parameters from the steps listed in Table 1 were used to build the model; these are primarily photo, etch, and wafer shape factors. Linear correlation with our response yielded

moderate correlations with the best result coming from the DRY STRIP SEM CD step (Figure 7).

Module Steps Used in Analysis
PHOTO REG
DRY STRIP SEM CD
BACKSIDE DEP SHAPE
PRE PHOTO SHAPE
PHOTO

Table 1: Steps Used in Model

Parameter	Pearson Correlation with DELTA_THK	Absolute Pearson Correlation with DELTA_THK
DELTA_OXIDE_THK_OXIDE	1.00	1.00
Cl2 Flow_MIN_DRY STRIP SEM CD	-0.52	0.52
Cl2 Flow_AVG_DRY STRIP SEM CD	-0.52	0.52
C4F6 Flow_MIN_DRY STRIP SEM CD	-0.52	0.52
FIELD X TRANS_PHOTO REG	0.51	0.51
C4F6 Flow_DRY STRIP SEM CD	-0.51	0.51
MaxInnerQMeritX_PHOTO REG	-0.49	0.49
C4F6 Flow_MAX_DRY STRIP SEM CD	-0.49	0.49
WARP_DEP SHAPE	0.49	0.49

Figure 7: Linear Correlation of Features with DELTA THK

As part of the exploratory analysis process, the distributions of features are examined. The distributions of the variables are analyzed to verify skewness or bimodality. A skewed distribution could yield low results for statistical tests. Skewed distributions can be handled by using the log, Box-Cox, or square root transformations. The spread of the data is then analyzed for data count, mean, and standard deviations. Next, some corrective action is done to clean the data, which includes correcting skewness, outlier removal, and imputation of missing values. Since sampling of wafers is usually not done across all process steps, due to time constraint, most steps lack overlap in wafer data measurements. In this case, features are filtered by how much data are available, so that the ones with fewer missing data points can be imputed to fill missing values.

After data exploration, correlation, data filtering, and imputation, there is a total of 1199 features left, which constitutes parameters belonging to the same steps. A 2-step feature selection process is used to narrow down features to the core ones that will be used for machine learning. The 2-step feature selection process includes elimination of features with high collinearity, then using an embedded feature selector to determine features of high importance to model.

Finally, several machine learning algorithms are attempted to model for the DELTA OXIDE THK from the selected upstream step features, after which an analysis is done to determine features having highest influence on the DELTA OXIDE THK response.

II. METHODOLOGY

A. Univariate analysis and Feature Distribution

The data set used has 1085 observations and 3805 parameter features across several upstream steps.
The steps with highest correlation coefficients with respect to response are shown in Figure 7.

Linear correlations with DELTA OXIDE THK do not show very high correlation coefficients which means the correlation is not very strong. The highest coefficient is 0.522 from the DRY STRIP SEM CD step. PHOTO REG and BACKSIDE DEP SHAPE are showing moderate correlation strength as well. Although correlation does not imply causation, it could be inferred that these steps have some significance to the outcome of the DELTA OXIDE THK measurements. This, however, should be confirmed with a multivariate analysis to determine the steps that hold highest importance in predicting DELTA OXIDE THK.

The distributions of these features are as shown in the correlation plots in Figure 8. The response itself shows a somewhat normal distribution (Figure 9).

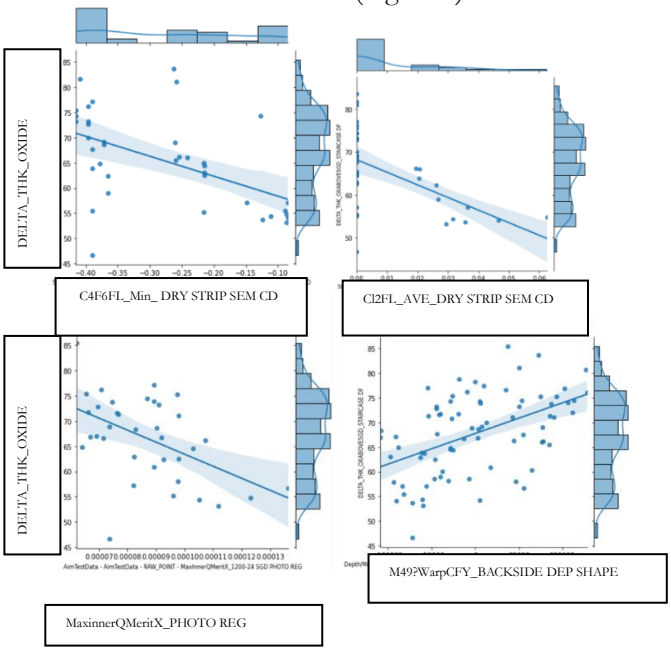

Figure 8: Regression Plots with Distributions of Highest Correlating Features with DELTA THK

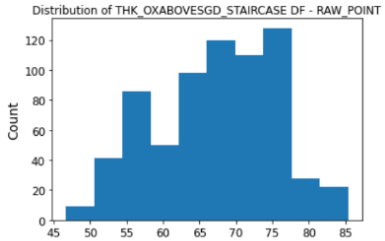

Figure 9: Distribution of DELTA OXIDE THK Response

B. *Data Cleaning and Imputation*

Since not all wafers are sampled across all steps, there is a lack of overlap in most step parameters, leading to missing values in rows across multiple steps. To correct this problem, an analysis on how much data is available per column and per row is done. Rows with no data present in any of the columns are dropped. Columns that have at least 20% data are kept. It is important to note that the typical practice is to keep columns with at least 70% of available data. This, however, poses a challenge as wafers are not sampled across all steps. Feature columns showing only one unique value are also dropped as correlation cannot be determined from a single value distribution. After filtering out rows and columns, the sample size drops to 692 rows and 1199 features.
Table 2 below shows a sample of available data for a few selected wafers.

Lot-Wafer	Step 1	Step 2	Step 3	Step 4	% Data Available per Row
X-01	X	X	X	X	100
Y-03			X	X	50
Z-02				X	25
W-01					0
% Data Available per Column	25	25	50	75	

Table 2: Sample Table of Data Availability and Overlap

Imputation of missing values is done using a KNN Imputer algorithm, which uses Lot Id, Wafer Id, and available inline data to estimate the missing data. Other methods of imputation include using column average, standard deviation, or modal value. KNN Imputer is preferred in this case as it considers other dimensions in the data.

Imputation of missing values is inferenced from majority data to fill the minority unavailable. In this case, there is a possibility of high bias in imputation as majority of data is unavailable.

III. RESULTS

A. *Feature Selection*

Feature selection is done in 2 steps prior to modeling. First, through elimination of features that highly correlate with each other. Some parameters, usually within the same step tend to be highly dependent on each other. They contribute less to the model and increase computational cost, which makes them not very useful for model building. These features are dropped based on correlation coefficient with each other. It is important to determine an appropriate threshold for the features as dropping too many may reduce prediction power of the model. This approach was used by M. Meiners and J. Franke in their ML application to sort out parameters showing fictitious correlations, with the use of a variance inflation factor (in our case, correlation coefficient threshold) (4).

Next, an embedded feature selector which uses the Random Forest machine learning algorithm is used to further narrow down features. The algorithm uses feature importance when fitting the data to model and can select features that hold high predicting power and output them.

The following table (Table 3) shows model performance with varying correlation coefficients used to drop highly collinear features, prior to using the embedded feature selector.

Good / OK / Bad	Correlation Coefficient Threshold	>=0.5	>=0.7	>=0.9
	# of Features Left	33	141	313
	# Features selected from Embedded Feature Selector	11	32	65
Train (553 observations)	Mean Absolute Error	0.25	0.22	0.03
	Percent Abs Error from Mean	0.38%	0.33%	0.05%
	Root Mean Square Error	0.79	0.77	0.16
	R-squared	0.99	0.99	1.0
	Adj R-squared	0.99	0.99	1.0
Test (139 observations)	Mean Absolute Error	1.7	1.62	1.48
	Percent Abs Error from Mean	2.51%	2.4%	2.18%
	Root Mean Square Error	3.76	3.36	3.4
	R-squared	0.80	0.84	0.83
	Adj R-squared	0.79	0.80	0.69

Table 3: Model Comparison with Varying Correlation Coefficient Thresholds

As seen in Table 3 above, majority of the features are highly correlated with each other, with the number of features dropping from 1199 to 331, when a correlation coefficient threshold of 0.9 is used. The Percent Absolute Error from Response Mean drops, and R-squared increases when more features are used. This, however, comes at the expense of the Adjusted R-squared which accounts for the dimension of the data and the number of observations being used. Since using a higher threshold of 0.9 causes a huge difference between R-squared and Adjusted R-squared and using a lower threshold of 0.5 decreases model performance; the threshold of 0.7 is used for feature selection. It not only has a higher Adjusted R-squared, but the difference from the R-squared value is not very large.

B. *Model Building*

Varying machine learning algorithms are tested and evaluated for model performance on the final data set. The key algorithms used are Random Forest, XGBoost Regressor, and Gradient Boost Trees. Other algorithms tested were; Kernel Ridge Regression, Lasso Regression, and SVR, but they yielded lower performance. Table 4 shows the performance across the three algorithms.

Good / OK / Bad	Metric	Random Forest	XGBoost Regressor	Gradient Boost Tree
	# of Features Used	34	34	34
Train (551 observations)	Mean Absolute Error	0.27	0.29	0.22
	Percent Abs Error from Mean	0.40%	0.44%	0.33%
	Root Mean Square Error	1.05	0.85	0.77
	R-squared	0.98	0.99	0.99
	Adj R-squared	0.98	0.99	0.99
Test (141 observations)	Mean Absolute Error	1.94	1.77	1.62
	Percent Abs Error from Mean	2.86%	2.62%	2.40%
	Root Mean Square Error	3.90	3.44	3.36
	R-squared	0.78	0.83	0.84
	Adj R-squared	0.71	0.78	0.80

Table 4: Score Card for Comparison of Model Algorithm Performance

The Gradient Boost Tree overall performs best on training data, showing lowest error and highest R-squared. The XGBoost performs least.

Hyperparameter tuning is essential for model building to optimize model performance. The data is split into train and test data in an 80:20 ratio. The train data has 553 rows and 34 feature columns, while the test has 139 rows and 34 feature columns. The rows here represent the wafers. A GridSearch algorithm is applied on the train data to find the best fit to the model based on hyperparameters such as max_depth, min_samples_leaf, min_samples_split, n_estimators, and random_state for consistent results. The best score outputted is 0.97 with a max_depth of 5, min_samples_leaf of 2, n_estimators of 120, and random_state of 10. This is then used to build the regression model.

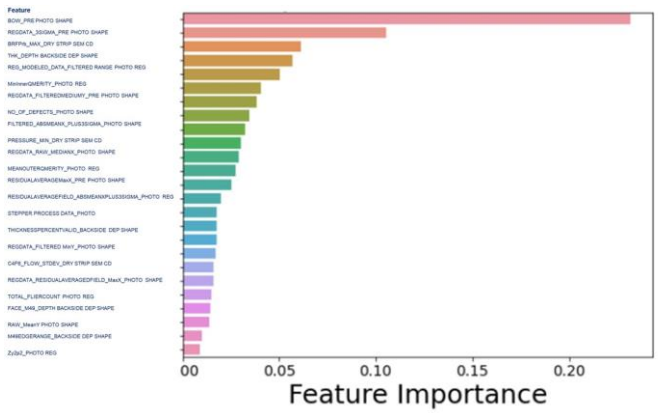

Figure 12: Feature Importance in Model Algorithm

The results from the model show a good correlation between the features used and DELTA OXIDE THK (Table 5). R-squared is 0.83 on test data which shows good goodness-to-fit. The adjusted R-squared is 0.78 which can be improved by further narrowing down the features used. The percent absolute error from mean is 2.44%. The threshold set by the SMEs to be a useful model is 10%. This model is very much below the threshold which is great for usability. There is a slight overfit between the train and test result as seen from the error and R-squared. The results overall look good on the test which makes it a decent model. Figures 10 and 11 below are references for model fit.

Performance Metric	Train	Test
Mean value of response	67.07	67.81
Standard Deviation from Mean of response	8.17	8.33
Mean value of prediction	67.07	68.09
Standard Deviation from Mean value of prediction	8.04	7.49
Mean Absolute Error of prediction	0.24	1.65
Percent absolute error from mean response	0.36%	2.44%
Root Mean Squared Error	0.76	3.39
R-squared	0.99	0.83
Adj R-squared	0.99	0.78

Table 4: Table of Results of Predictions with GBT Algorithm

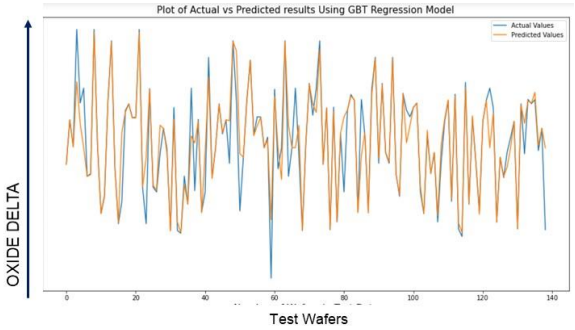

Figure 10: Predictions vs Actual

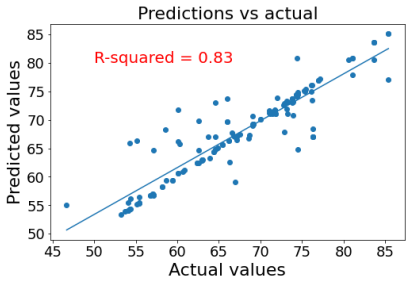

Figure 11: Fitted Scatter Plot of Predictions vs Actual

Further tuning of the model can be done to improve it and eliminate overfit. The highest feature importance to the model is from the PRE PHOTO SHAPE followed by DRY STRIP SEM CD (Figure 12).

It was anticipated that variation could be coming from SHAPE features such as BOW. Correlation plots of top feature importance is shown in Figure 13.

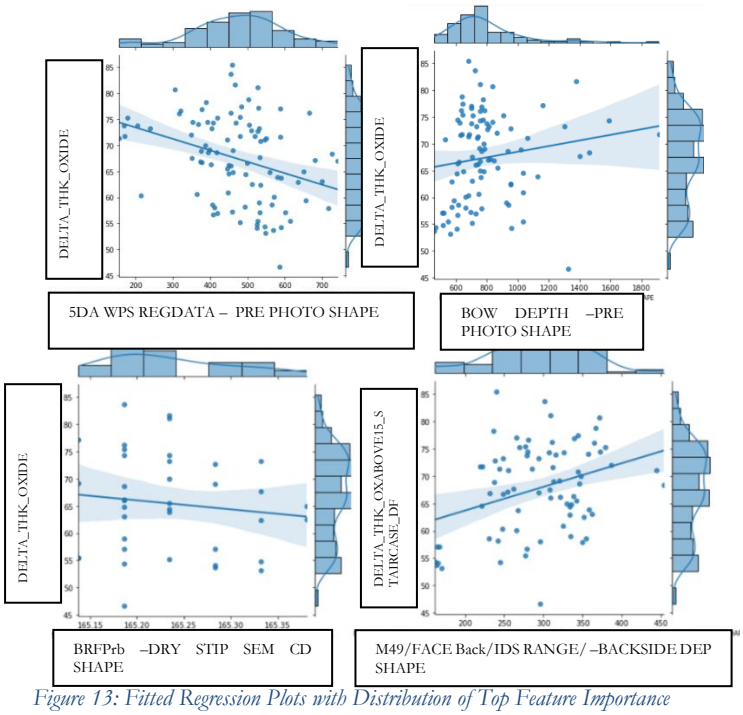

Figure 13: Fitted Regression Plots with Distribution of Top Feature Importance

The majority of the data points are clustered in lower values of the PRE PHOTO SHAPE BOW parameter whilst there is a slight negative relationship in the STRIP SEM CD parameter as the parameter values increase.

The model results above are produced from 34 features. Further narrowing down the features to 12 shows very similar results to using 324 features with only a slight drop in r-squared from 0.83 to 0.81 (Figure 14), and slight drop in error (Table 6). The top features come from PRE PHOTO SHAPE, BACKSIDE DEP SHAPE, PHOTO REG and DRY STRIP SEM CD. Narrowing down features further showed a further drop in r-squared and an increase in error, hence the minimum number of features used was left at 12. An increase in number of features from 34 features, as shown in the Table 3 yielded a smaller adjusted r-squared as the adjusted r-squared is dependent on size of data and dimension. The ideal is to obtain optimum results with least number of features.

Performance Metric	Train	Test
Mean value of response	67.07	67.81
Standard Deviation from Mean of response	8.17	8.33
Mean value of prediction	67.07	68.22
Standard Deviation from Mean value of prediction	8.04	7.44
Mean Absolute Error of prediction	0.24	1.72
Percent absolute error from mean response	0.43%	2.54%
Root Mean Squared Error	0.84	3.58
R-squared	0.99	0.81
Adj R-squared	0.99	0.80

Table 5: Table of Model Performance with Top 12 Feature Importance

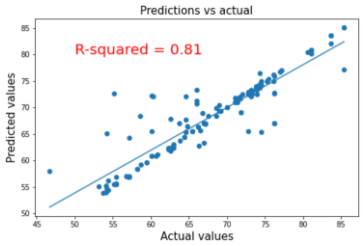

Figure 14: Fitted Scatter Plot of Predicted vs Actual with Top 12 Feature Importance

Further analysis is done to evaluate how PRE PHOTO SHAPE on its own affects the outcome of the DELTA OXIDE THK. Linear correlations show LSC_Data_Mean, Bow and Warp features have moderate linear correlation with the DELTA OXIDE THK (Figure 15).

Parameter	Correlation with DELTA_OXIDE_THK	Absolute Correlation with DELTA_THK_OXIDE
DELTA_OXIDE_THK	1.000	1.000
LSCDATA MEAN	0.412	0.412
MEANLSC_PRE PHOTO SHAPE	0.410	0.410
WARP_CFXY_PRE PHOTO SHAPE	0.410	0.410
BOW_X_DF_PRE PHOTO SHAPE	0.410	0.410
BOW_X_RAW_PRE PHOTO SHAPE	0.410	0.410
WARP_CFX_PRE PHOTO SHAPE	0.410	0.410
WARP_CFXY_PRE PHOTO SHAPE	0.410	0.410
WARP_BF_PRE PHOTO SHAPE	-0.408	0.408
BOW_DF PRE PHOTO SHAPE	0.406	0.406
BOW_BF PRE PHOTO SHAPE	0.405	0.405

Figure 15: Linear Correlation of PRE PHOTO SHAPE Features with DELTA OXIDE THK

The 261 parameters from PRE PHOTO SHAPE are passed through the feature selection process described earlier, identifying 11 features from the step used for model building with results shown in Table 7.

Performance Metric	Train	Test
Mean value of response	66.62	70.58
Standard Deviation from Mean of response	8.19	7.18
Mean value of prediction	66.62	66.75
Standard Deviation from Mean value of prediction	3.7	2.68
Mean Absolute Error of prediction	4.17	5.82
Percent absolute error from mean response	6.27%	8.25%
Root Mean Squared Error	5.04	7.57
R-squared	0.62	-0.11
Adj R-squared	0.57	-1.11

Table 6: Table of Model Results with PRE PHOTO SHAPE

The results show decent ability to predict response with error percent at 8.25% which is within the 10% cut off limit. The R-squared, however, is weak. SHAPE alone could be a good contributor to the DELTA OXIDE THK but the drop in performance is an indicator that several factors affect the uniformity variation and so SHAPE alone, though is within reasonable error range, will need to be combined with other good predictors in order to show stronger correlation. The 11 features used in PRE PHOTO SHAPE and their feature importance is shown in Figures 16 and 17.

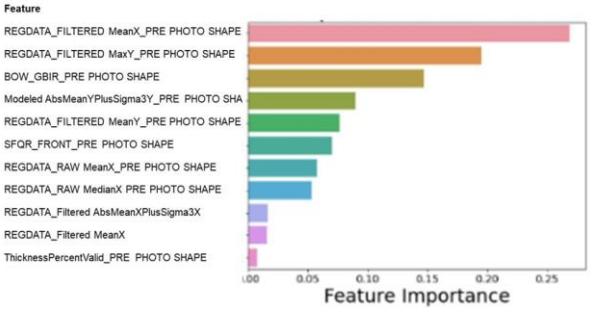

Figure 16: Feature Predictors from PRE PHOTO SHAPE and their Feature Importance

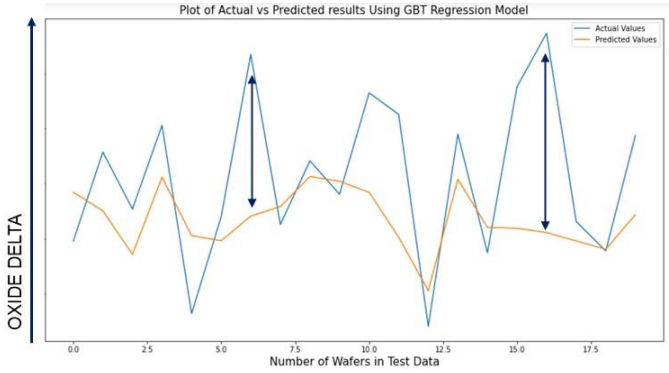

Figure 17: Graph of Predicted vs Actual Results with PRE PHOTO SHAPE

IV. CONCLUSION AND FUTURE WORK

Although this report mainly discusses modeling using upstream parameters, the DELTA OXIDE THK response in POLY CMP is affected by several additional factors, including Recipe Process Adjustment (RPA) parameters at that step. Modeling with upstream parameters pointed to the PRE SHAPE as the highest feature of importance. However, the response is affected by multiple steps along the process line as shown in Figure 12. Model results from using only SHAPE parameters did not yield good performance as shown in Table 7 and Figure 17. The R-squared was weak and percent error was much higher than in the multi-step model, despite being within specification limit of 10% error. The multi-step model, however, had a good R-squared of 0.83 and percent error of 2.44%, which were well within specification limits, making it a better approach than using a single upstream step. Further validation of the model on new data will be essential to testing how accurate our model is in predicting new DELTA OXIDE THK on new test samples being produced. The next step in this project is to determine implementation of a combined model, using both the upstream step model and the RPA model to predict the outcome of DELTA OXIDE THK. This could be achieved with a usability ratio, with higher importance given to the process model as this is directly connected to the response and has a higher correlation to it.

The upstream steps model will serve as a feedforward controller while the RPA model will serve as a feedback controller. Also, an understanding on how to implement a R2R model from the combination of RPA and upstream step models is needed with suggestions from the CMP process owners and subject matter experts.

V. REFERENCES

[1] J. Yu and P. Guo, "Run-to-Run Control of Chemical Mechanical Polishing Process Based on Deep Learning Reinforcement Learning", *IEEE Transactions on Semiconductor Manufacturing,* Vol. 33, No. 3, August 2020.

[2] M.-H. Hsu, C.-C Lin, H.-M. Yu, K.-W. Chen, T. Luoh, L.-W. Yang, T.-H. Yang and K.-C. Chen, "Advanced CMP Process Control by using Machine Learning Image Analysis" *2021 IEEE Interconnect Technology Conference (IITC)* Online July 6-9, 2021.

[3] R. Ghulghazaryan and J. Wilson, "Application of Machine Learning and Neural Networks for Generation of Pre-CMP Profiles of Advanced Deposition Process for CMP Modeling" *ICPT 2017,* October 2017, Leuven, Belgium.

[4] M. Moritz and J. Franke, "Concept of a Machine Learning supported Cross-Machine Control Loop in the Ramp-Up of Large Series Manufacturing." *in 2020 IEEE 11th International Conference on Mechanical and Intelligent Manufacturing Technologies*, Cape Town, 2020.

VI. ACKNOWLEDGEMENTS

The authors would want to acknowledge Bindu Muthangi for pulling data and formatting the data that was used in the analysis.